

Building CoronaWhy Knowledge Graph

FREYA Guest webinar

Slava Tykhonov
Senior Information Scientist
(DANS-KNAW, the Netherlands)

01.09.2020

About me: DANS-KNAW projects (2016-2020)

- CLARIAH+ (ongoing)
- EOSC Synergy (ongoing)
- SSHOC Dataverse (ongoing)
- CESSDA DataverseEU 2018
- Time Machine Europe Supervisor at DANS-KNAW
- PARTHENOS Horizon 2020
- CESSDA PID (Personal Identifiers) Horizon 2020
- CLARIAH
- RDA (Research Data Alliance) PITTS Horizon 2020
- CESSDA SaW H2020-EU.1.4.1.1 Horizon 2020



Source: [LinkedIn](#)

Motivation



Slava Tykhonov 4:05 PM

Born in Kyiv, Ukraine but raised as a scientist in the Netherlands, during my career I saw a lot of cases where people refused to collaborate and work together due to own ambitions and wrong vision. I truly believe that dangerous things like coronavirus are possible in the modern society only because people are competing against each other and don't want to share their knowledge and competence, and work together in order to find a solution for a problem.

Stupidity, ignorance and limitness has no nationality, it's just a common thing that killing this world and COVID-19 is just one of the challenges. I'm here to open everything that should be open for the humanity, build the collaboration between people, speed up the innovation and start the development of the research infrastructure that will allow to bring Science back to the policy table, and quickly respond to the current and future challenges. I would say, we should be prepared for the technological future and don't afraid to disrupt the world. (edited)



6



1



3



7 weeks in lockdown in Spain



[Resistere \(I will resist\)](#)

About CoronaWhy

www.coronawhy.org



[Home](#)

[Daily Progress](#)

[Calendar](#)

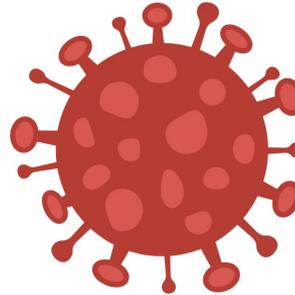
[JOIN THE FIGHT!](#)

FIGHTING CORONAVIRUS WITH ARTIFICIAL INTELLIGENCE

We are a globally distributed, volunteer-powered research organisation, trying to assist the medical community's ability to answer key questions related to COVID-19

[JOIN THE FIGHT](#)

[LEARN MORE](#)



Who we are?

Artur Kiulian started CoronaWhy because he realized that we are all in this together. Now CoronaWhy is an international group of 900+ volunteers whose mission is to improve global coordination and analysis of all available data pertinent to the COVID-19 outbreak and ensure all findings reach those who need them.

It's impossible to list everyone out but we will eventually.

1300+ people registered in the organization, more than 300 actively contributing!

COVID-19 Open Research Dataset Challenge (CORD-19)

It's all started from [this](#) (March, 2020):

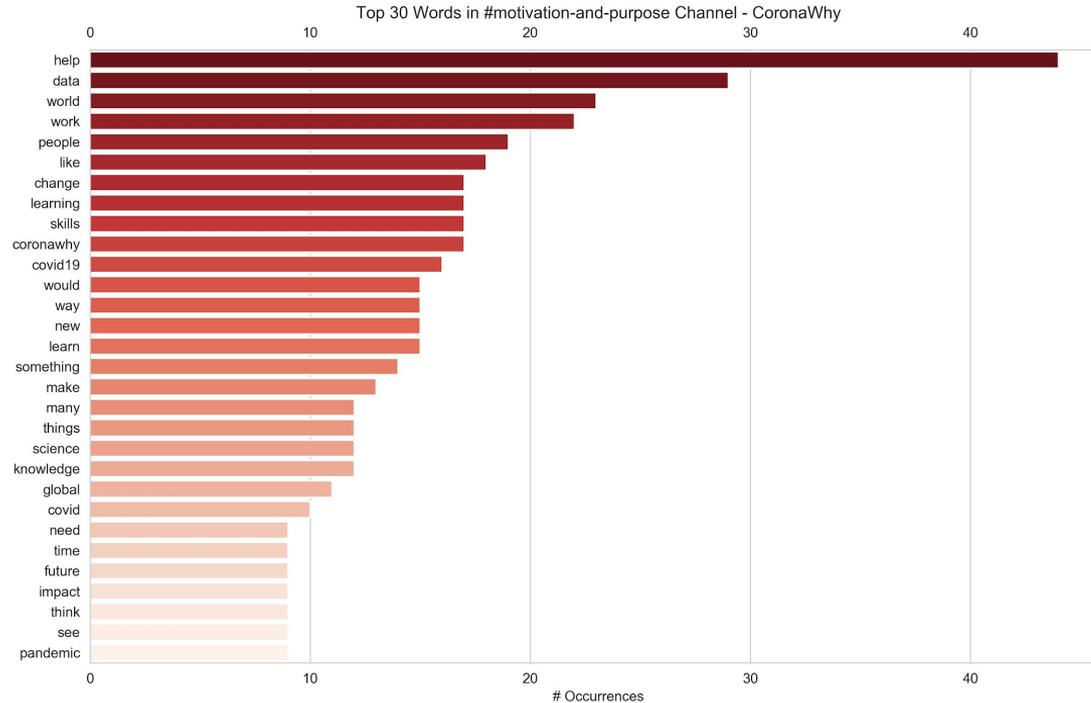
“In response to the COVID-19 pandemic and with the view to boost research, the Allen Institute for AI together with CZI, MSR, Georgetown, NIH & The White House is collecting and making available for free the COVID-19 Open Research Dataset (CORD-19). This resource is updated weekly and contains over 52,000 scholarly articles, including 41,000 with full text, about COVID-19 and other viruses of the coronavirus family.” ([Kaggle](#))

Collaborators

CORD-19 was made possible by the [Semantic Scholar team](#) at the [Allen Institute for AI](#) in collaboration with the following institutions:



Motivation of CoronaWhy community members



Credits: Andre Ye

CoronaWhy Funding

Initial: \$5k from Google on GCP and \$4k from Amazon on AWS (April 2020)

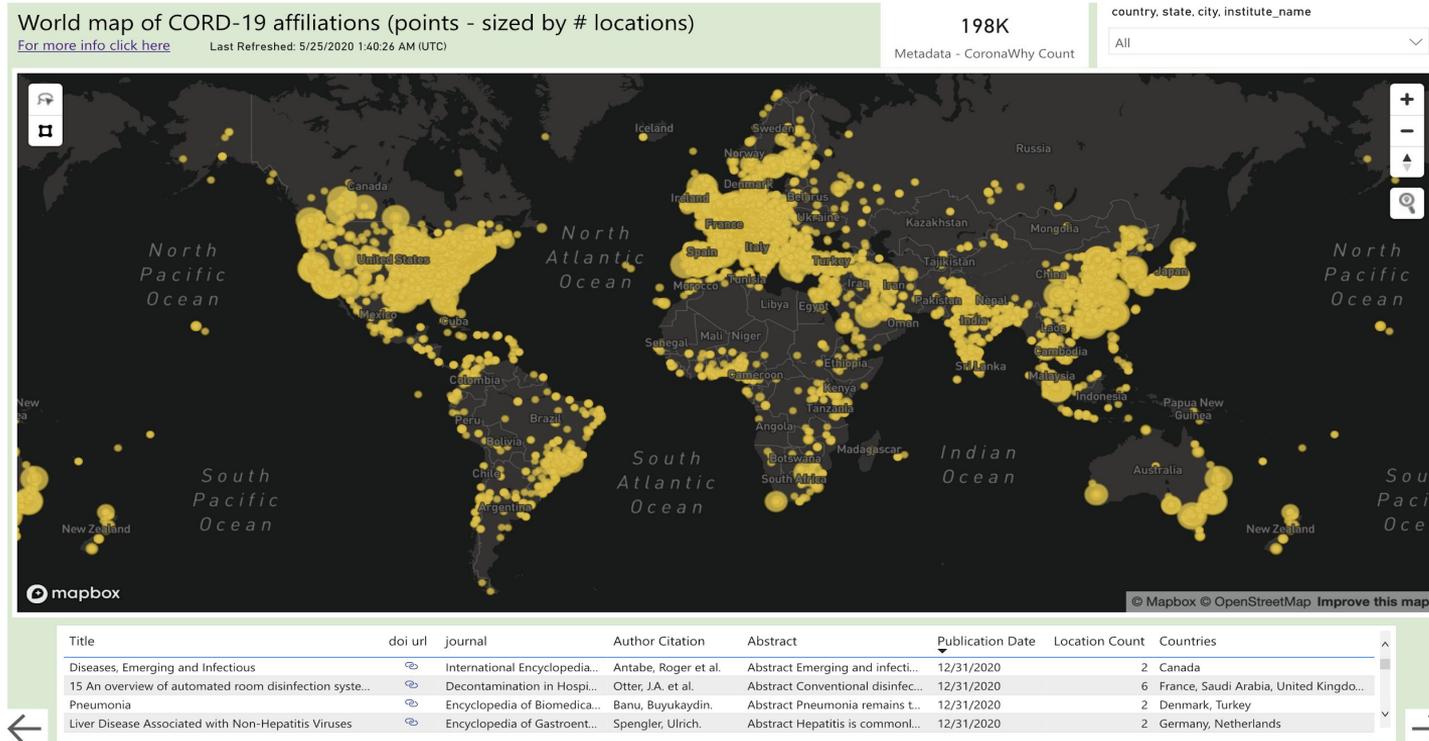
Donations: \$9k and 15k british pounds to sustain CoronaWhy infrastructure

 High power Virtual Machine for Data Visualization	 Our first-ever server	 Google Cloud \$5000 credit for Google Cloud Platform	 Business class plan
 Standard plan	 1 year business site-plan	 Free Virtual Classrooms	 Pro Bono Public Relations
 Security Staff	 Pro plan	 Collaborative ecosystem for complex local and global issues.	 Collective Intelligence Partners

CoronaWhy Community Tasks (March-April)

1. [Task-Risk](#) helps to identify risk factors that can increase the chance of being infected, or affects the severity or the survival outcome of the infection
2. [Task-Ties](#) to explore transmission, incubation and environment stability
3. [Match Clinical Trials](#) allows exploration of the results from the [COVID-19 International Clinical Trials](#) dataset
4. [COVID-19 Literature Visualization](#) helps to explore the data behind the AI-powered literature review
5. [Named Entity Recognition](#) across the entire corpus of COVID-19 papers with full text

CORD-19 affiliations recognized with Deep Learning



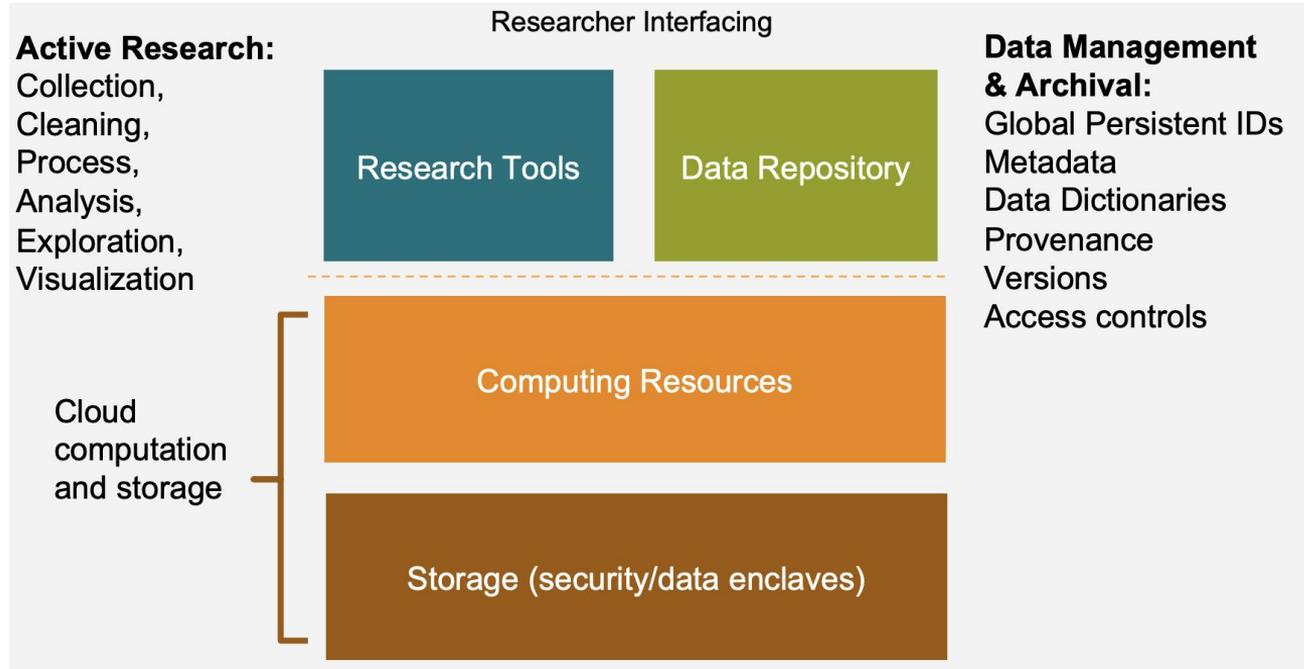
Source: [CORD-19 map visualization](#) and [institution affiliation data](#)

Collaboration with other organizations

- Harvard Medical School, INDRA integration
- Helix Group, Stanford University
- NASA JPL, COVID-19 knowledge graph and GeoParser
- Kaggle, coronamed application
- Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS, knowledge graph
- dcyphr, a platform for creating and engaging with distillations of academic articles
- CAMARADES (Collaborative Approach to Meta-Analysis and Review of Animal Data from Experimental Studies)

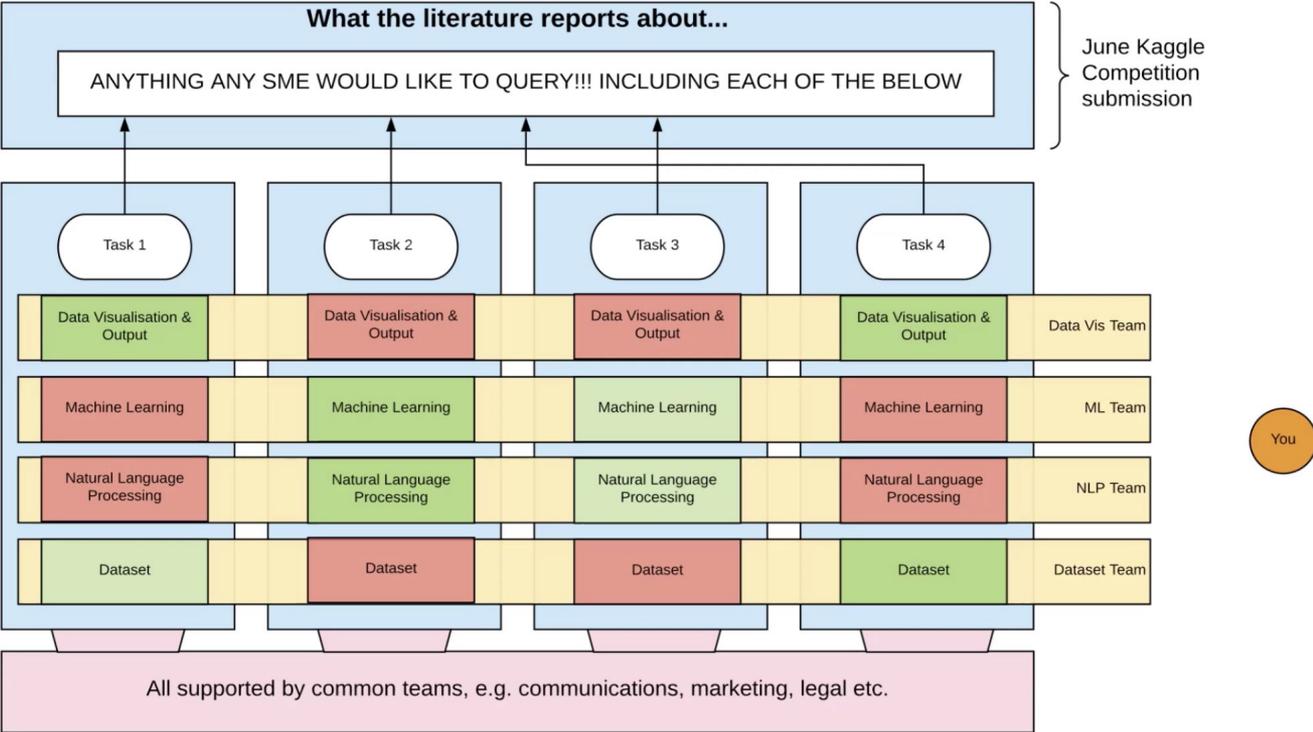
We've got almost endless data streams...

Looking for Commons



Merce Crosas, "[Harvard Data Commons](#)"

Building a horizontal platform to serve vertical teams



Source: [CoronaWhy infrastructure](#) introduction

Turning FAIR into reality!

WHAT IS FAIR ?

Findable:

- F1 (meta)data are assigned a globally **unique** and **persistent** identifier; **FM-F1A FM-F1B**
- F2 data are described with **rich metadata**; **FM-F2**
- F3 metadata clearly and explicitly include the **identifier of the data** it describes; **FM-F3**
- F4 (meta)data are registered or **indexed** in a searchable resource; **FM-F4**

Interoperable:

- I1 (meta)data use a formal, accessible, shared, and broadly applicable **language for knowledge representation**. **FM-I1**
- I2 (meta)data use **vocabularies that follow FAIR principles**; **FM-I2**
- I3 (meta)data include **qualified references** to other (meta)data; **FM-I3**

Accessible:

- A1 (meta)data are retrievable by their identifier using a standardized communications protocol;
 - A1.1 the protocol is **open, free, and universally implementable**; **FM-A1.1**
 - A1.2 the protocol allows for an **authentication and authorization** procedure, where necessary; **FM-A1.2**
- A2 metadata are accessible, **even when the data are no longer available**; **FM-A2**

Reusable:

- R1 meta(data) are richly described with a plurality of accurate and relevant attributes;
 - R1.1 (meta)data are released with a clear and **accessible data usage license**; **FM-R1.1**
 - R1.2 (meta)data are associated with **detailed provenance**; **FM-R1.2**
 - R1.3 (meta)data meet domain-relevant **community standards**; **FM-R1.3**

Sci. Data 3:160018 doi: 10.1038/sdata.2016.18 (2016)
<http://fairmetrics.org>
<https://github.com/FAIRMetrics/Metrics/blob/master/ALL.pdf>



Standing on the Shoulders of Giants: infrastructure



Type of action & funding:
Research and Innovation action
(INFRAEOSC-04-2018)

Partners: 47
(20 beneficiaries + 27 LTPs)

SSH ESFRI Landmarks and Projects
& international SSH data infrastructures

Project budget:
€ 14,455,594.08

Duration: 40 months
(January 2019 – 30 April 2022)

Project website:
www.SSHOpenCloud.eu



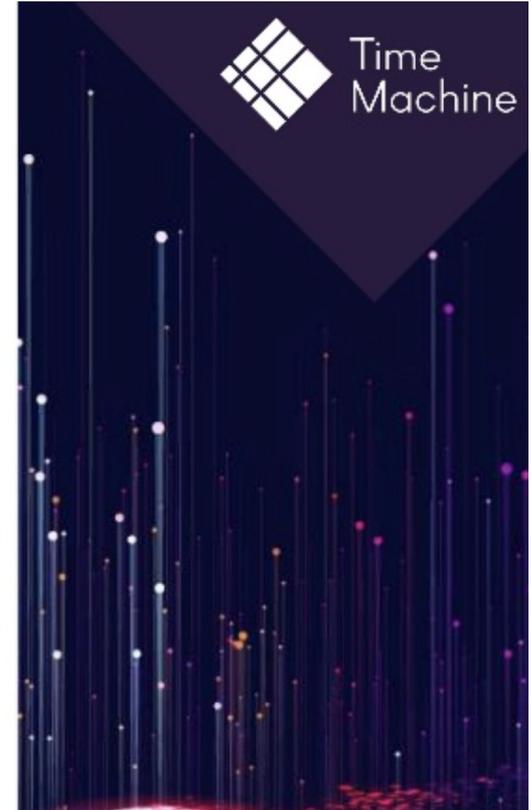
Objectives:

- creating the social sciences and humanities (SSH) part of European Open Science Cloud (EOSC)
- maximising **re-use** through **Open Science** and **FAIR** principles (standards, common catalogue, access control, semantic techniques, training)
- interconnecting existing and new infrastructures (clustered cloud infrastructure)
- establishing appropriate **governance model** for SSH-EOSC

Standing on the Shoulders of Giants: Big Data of the Past

Time Machine is ...

- An international **collaboration** to bring 5000 years of European history to life
- Digitising millions of **historical documents**, painting and monuments
- The **largest** computer simulation ever developed
- An **open access**, interactive resource



Dataverse as data integration point

Dataverse Search User Guide Support Sign Up Log In

COVID-19 Data Hub

Metrics 3,429 Downloads [Contact](#) [Share](#)

Information and Data hub produced by all [CoronaWhy](#) research groups. Please [join us](#) if you want to help in the fight against COVID-19.
Disclaimer: at the moment all materials published on this site are available for public for the demonstration purposes, without [DOI Persistent Identifiers](#).

CoronaWhy Task Ties | Pandemics in History | INDRA | COVID-19 Knowledge Graphs

Search this dataverse... [Find](#) [Advanced Search](#)

Dataverses (24)
 Datasets (506)
 Files (22,755)

Dataverse Category
Research Group (10)
Laboratory (8)
Research Project (4)
Organization or Institution (1)

Publication Year
2020 (532)

Author Name
CoronaWhy Labs (9)
Tykhonov, Vyacheslav (7)
Chen, Christine (4)
CoronaWhy (3)
Lofgran, Alex (3)

1 to 10 of 532 Results [Sort](#)

NASA JPL Knowledge Graph extracted from COR-19 collection
Aug 24, 2020 - NASA Data Hub
McGibbney, Lewis John, 2020, "NASA JPL Knowledge Graph extracted from COR-19 collection", <https://doi.org/10.5072/FK2-OLXCRD>, COVID-19 Data Hub, V1
This dataset contains a knowledge graph from the COVID-19 Open Research Dataset (CORD-19) dataset. File covid19_knowledge_graph.ttl can be loaded into Apache Jena's Fuseki server (or any other SPARQL server which permits ingest of TTL RDF graphs).

NASA Data Hub (NASA)
Aug 24, 2020
NASA (National Aeronautics and Space Administration) collaborative efforts together with CoronaWhy community.

INDRA statements
Aug 24, 2020 - INDRA
Gyori, Benjamin M., 2020, "INDRA statements", <https://doi.org/10.5072/FK2-NQAUNA>, COVID-19 Data Hub, V1
INDRA (the Integrated Network and Dynamical Reasoning Assembler) assembles information about biochemical mechanisms into a common format that can be used to build several different kinds of explanatory models. This dataset contains mechanistic information from multiple sources is

- Available as a service for the community from April, 2020
- Used by CoronaWhy vertical teams for the data exchange and share
- Intended to help researchers to make their data FAIR
- One of the biggest COVID-19 data archives in the world with 700k files
- New teams are getting own data containers and can reuse data collected and produced by others

<http://datasets.coronawhy.org>

Dataset from CoronaWhy vertical teams

Dataverse Search User Guide Support Sign Up Log In

CoronaWhy Task Risk (CoronaWhy)

COVID-19 Data Hub > CoronaWhy Task Risk >

COVID-19 risk factors

Version 1.0

 Mayya Lihovodov; Pranjalata Tiwari; Ansun Sujoe; Guillermo Blanco; Iason Konstantinidis; Kriti Mahajan; Robbie Edwards; Vijay Datta; Michael Wang; Lukasz Gagala; Brandon Eychaner; Mohammad Tanweer; Anrew Wood; Kevin Lee; Samtha Reddy; Mark Koranda; Ruslan Olinyk, MD; Mike Honey; Randall Brown, MD; Artur Kiulian, 2020, "COVID-19 risk factors", <https://doi.org/10.5072/FK2/3OZLV6>, COVID-19 Data Hub, V1

[Cite Dataset](#) [Learn about Data Citation Standards.](#)

[Access Dataset](#)
[Contact Owner](#) [Share](#)

Dataset Metrics [?](#)
184 Downloads [?](#)

Description [?](#)

A major topic of interest among researchers is the study of the various risk factors related to COVID-19. A risk factor is anything that increases the chance of being infected, or affects the severity or the survival outcome of the infection. Many of the papers in the dataset are studies on the severity and outcome of the infection, without, however, any systematic documentation that would be easily searchable.

The focus of this study is to extract and present in a meaningful and easily accessible way scientific papers that are related to risk factors associated with viral diseases through a procedure that can be automated as much as possible.

At the current stage, a semi-automated approach is implemented using manual review of retrieved papers. It is important to note that through the proposed procedure a small subset of papers is manually reviewed, the ones that are identified as most probable to be relevant to a specific risk factor. This brings the volume of papers for review down to less than 100-200 instead of multiple thousands, rendering the review task feasible in much shorter timeframes.

Also, at the current stage the paper extraction is limited to the following factors:

Environmental: Pollution, Population Density, Humidity, Temperature
Comorbidity: Heart diseases
Demographics: Senior age
Lifestyle: Smoking

The above risk factors were identified as being the most important by the medical community. An extensive list of risk factors is provided under section 4 below and is subject of a future extension of this study.

Source: [CoronaWhy Dataverse](#)

COVID-19 data files verification

The screenshot shows the Dataverse interface for a dataset titled 'coronavirus'. The subject is 'Medicine, Health and Life Sciences'. The interface includes tabs for 'Files', 'Metadata', 'Terms', and 'Versions'. There are options to 'Change View' between 'Table' and 'Tree' views. A search bar is present with the text 'Search this dataset...'. Below the search bar, there are filters for 'File Type: All', 'Access: All', and 'File Tag: All'. A 'Sort' button is also visible. The main content area displays a list of files, with the first three visible:

- age-distribution-died-and-survivors.xlsx**
data/hospitalized/age-distribution-died-and-survivors.xlsx/
MS Excel Spreadsheet - 5.0 KB - May 30, 2020 - 0 Downloads
MD5: 7343b296eaf59c98fc426e1a03234221
Data snapshot from https://raw.githubusercontent.com/Sikerdebaard/dutchcovid19data/master/data/hospitalized/age-distribution-died-and-survivors.xlsx
Metadata tags: patients_deceased, dutchcovid19data, age_group, patients_recovered
- age-distribution-status.xlsx**
data/hospitalized/age-distribution-status.xlsx/
MS Excel Spreadsheet - 5.1 KB - May 30, 2020 - 0 Downloads
MD5: ea9604d071e3597748ec185b19d956cc
Data snapshot from https://raw.githubusercontent.com/Sikerdebaard/dutchcovid19data/master/data/hospitalized/age-distribution-status.xlsx
Metadata tags: patients_deceased, dutchcovid19data, patients_in_hospital, age_group, patients_recovered
- age-distribution-status.xlsx**
data/age-distribution-status.xlsx/
MS Excel Spreadsheet - 5.1 KB - May 30, 2020 - 0 Downloads
MD5: 9c418de86b10f0aef872200885b5a4a3
Data snapshot from https://raw.githubusercontent.com/Sikerdebaard/dutchcovid19data/master/data/age-distribution-status.xlsx
Metadata tags: dutchcovid19data, patients_in_hospital, recovered_patients, patients_in_icu, deceased_patients, age_group

We do a verification of every file by importing its contents to dataframe.

All column names (variables) extracted from tabular data available as labels in files metadata

We've enabled Dataverse data previewers to browse through the content of files without download!

We're starting internal challenges to build ML models for the metadata classification

Dataverse content in Jupyter notebooks

Dataverse content, data and metadata import to Pandas Dataframes ☆

File Edit View Insert Runtime Tools Help [All changes saved](#)

+ Code + Text

RAM Disk Editing

Now read the content of first file to dataframe

```
FILEID=pdfiles.loc['id'][0]
fileURL = "%s/api/access/datafile/%s" % (BASE_URL, FILEID)
df = pd.read_csv(fileURL)
df.head()
```

Unnamed: 0	Risk Factor	Title	Keyword/Ngram	No of keyword occurrence in Paper	paper_id	URL	Sentences	Authors	Correlation	Design Methodology	s	
0	0	age	Coronavirus-like particles in nonhuman primate...	[older age group]	1	3cf9a172522a7db0df9e436029707bb6e3e0ff8c	[https://www.ncbi.nlm.nih.gov/pmc/articles/PM...]	['It might be assumed that coronaviruses are n...]	['Smith, G. C.; Lester, T. L.; Heberling, R. L...]	It might be assumed that coronaviruses are not...	NaN	
1	1	age	Estimates of the severity of coronavirus disea...	['60 years and over', 'older age group']	7	ac2a1ba62fdf52eb276b42b22ed3d927b5330b1	[https://doi.org/10.1016/s1473-3099(20)30243-...]	['Reported cases in Wuhan were more frequent i...]	['Verity, Robert; Okell, Lucy C; Dorigatti, Il...]	Reported cases in Wuhan were more frequent in ...	In cases reported outside of mainland China, ...	nur obser in 1
2	2	age	Infections in travellers returning to Turkey f...	['65 years old']	3	f33e3be8c6ec1d348cf8983037dcf8adb25e7f94	[https://www.ncbi.nlm.nih.gov/pmc/articles/PM...]	['Seventy four (40 %) of them were ≥ 65 years ...]	['Erdem, H.; Ak, O.; Elaldi, N.; Demirdal, T.;...]	A total of 185 Turkish patients were recruited...	NaN	
3	3	age	The use of corticosteroid as treatment in SARS...	['patients older than']	1	52f5440ec7a22706f95be3c6f5e0ed2e940e4945	[https://doi.org/10.1016/j.jinf.2004.09.008',...]	['A total of 80 patients older than 18 years o...]	['Auyeung, Tung Wai; Lee, Jenny S.W.; Lai, Win...]	NaN	NaN	
4	4	age	Burden, seasonal pattern and symptomatology of...	[older age group]	5	0957f96f8188f65cc145464dc7882abd259e0f5f	[https://doi.org/10.1016/j.cmi.2015.05.027', ...]	['On comparison of the two age groups , viral...]	['Wei, L.; Chan, K.-H.; Ip, D.K.M.; Fang, V.J....]	On comparison of the two age groups , viral a...	NaN	

Source: [Dataverse examples on Google Colabs](#)

COVID-19 Data Crowdsourcing

Dataset in CoronaWhy Dataverse #579

New issue

 Open k-goncharova opened this issue 2 days ago · 4 comments



k-goncharova commented 2 days ago



Hello,Your dataset was added to CoronaWhy (<https://www.coronawhy.org/>) Data Lake on Dataverse as a piece of common COVID-19 dataframe <https://datasets.coronawhy.org/dataset.xhtml?persistentId=doi:10.5072/FK2/A20BEO>

Would you be willing to help with maintenance of your dataset in Dataverse, e.g. adding the relevant metadata and keeping the dataset up-to-date? That will help to make the dataset findable and accessible for medical science community.

Assignees

No one assigned

Labels

None yet

Projects

None yet

Milestone

No milestone

Linked pull requests

Successfully merging a pull request may close this issue.

None yet

Notifications

Customize

 Subscribe

You're not receiving notifications from this thread.

2 participants



swsoyee commented 2 days ago

Owner



Hi, @k-goncharova

Sure, I will keep updating my dataset, and what should I do in Dataverse?



k-goncharova commented 2 days ago

Author



Great, thank you. Please register in CoronaWhy Dataverse <https://datasets.coronawhy.org/> and I'll give you the Dataset creator permissions to your dataset. Then please add relevant metadata - description and keywords to your dataset.



swsoyee commented 2 days ago

Owner



Okay, my user's name is swsoyee .

I'm a little busy these days, so the progress maybe slow, please forgive me.

CoronaWhy data management team does the review of all harvested datasets and try to identify the important data.

We're approaching github owners by creating issues in their repos and inviting them to help us.

More than 20% of data owners joining CoronaWhy community or interested to curate their datasets.

Bottom-up data collection works!

Challenge of data integration and various ontologies

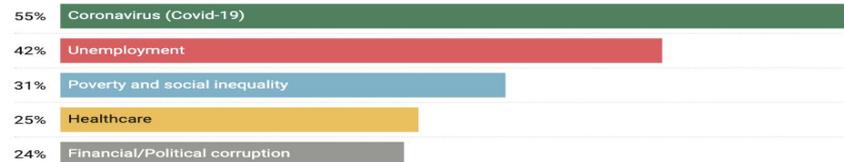
CORD-19 collection processing with NLP pipeline:

- manual **annotation** and **labelling** of COVID-19 related papers
- automatic **entity extraction** and **classification** of text fragments
- **statements** extraction and curation
- linking papers to specific research questions with **relationships** extraction

Dataverse Data Lake streaming COVID-19 datasets from various sources:

- medical data
- socio-economic Data
- political data and statistics

The top five global concerns



Research among adults aged 16-64 in 27 participating countries. c. 19,000 per month. (May 2020).
Source: Global Advisor · [Get the data](#) · Created with [Datawrapper](#)

The importance of standards and ontologies

Generic controlled vocabularies to link metadata in the bibliographic collections are well known: ORCID, GRID, GeoNames, Getty

Medical knowledge graphs powered by:

- Biological Expression Language (BEL)
- Medical Subject Headings (MeSH®) by U.S. National Library of Medicine (NIH)
- Wikidata (Open ontology) - Wikipedia

Integration based on metadata standards:

- MARC21, Dublin Core (DC), Data Documentation Initiative (DDI)

Biological Expression Language (BEL)



BEL Commons 3.0 Preview

An environment for **curating**, **validating**, and **exploring** knowledge assemblies encoded in Biological Expression Language (BEL) to support **elucidating** disease-specific, mechanistic insight.

Catalog

View summaries and statistics over curated networks and as a first step towards exploration and visualization.

[Networks](#)[Nodes](#)[Edges](#)[Citations](#)

Query

Build a network by investigating the knowledge related to interesting biological entities, chemical matter, authors, or publications.

[Build Query](#)

Terminologies

View the underlying namespaces and annotations along with their respective uniform resource locators and version numbers.

[Resources 29](#)

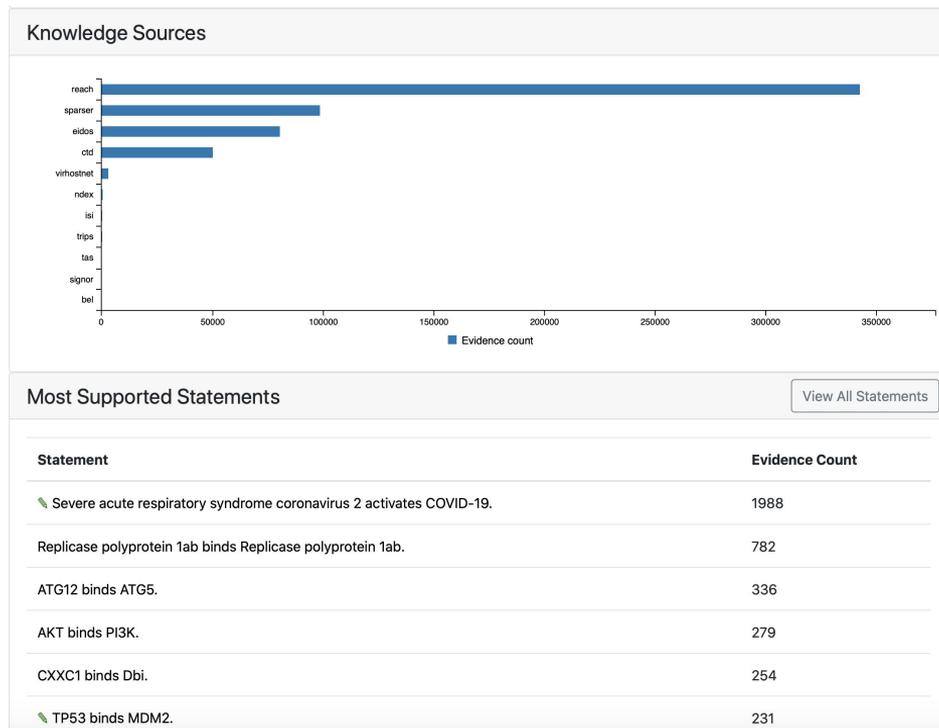
About

BEL Commons is a free, open-source web platform for Biological Expression Language built on top of [PyBEL](#).

If you find BEL Commons useful in your work, please consider citing: Hoyt, C. T., Domingo-Fernández, D., & Hofmann-Apitius, M. (2018). [BEL Commons: an environment for exploration and analysis of networks encoded in Biological Expression Language](#). *Database*, 2018(3), 1–11.

BEL was integrated in CoronaWhy infrastructure in April, 2020

Statements extraction with INDRA



“INDRA (Integrated Network and Dynamical Reasoning Assembler) is an automated model assembly system, originally developed for molecular systems biology and currently being generalized to other domains.”

Developed as a part of Harvard Program in Therapeutic Science and the Laboratory of Systems Pharmacology at Harvard Medical School.

<http://indra.bio>

Knowledge Graph curation in INDRA

	Effector activates transcription, DNA-templated.	0/50	JSON
	Phenobarbital increases the amount of CYP3A4.	0/50	JSON
	Oseltamivir inhibits Influenza, Human.	0/50	JSON
	Arg-Val activates Asthma.	49/49	JSON
reach	It remains unknown if RV induces the development of wheeze and asthma or if asthmatics are more susceptible to RV infection.	18234348	
reach	Figure 3: A mouse model of RV-induced asthma exacerbation.	24278777	
	<div style="border: 1px solid black; padding: 5px;"><ul style="list-style-type: none">✓ CorrectEntity BoundariesGroundingNo RelationWrong RelationActivity vs. AmountPolarityNegative ResultHypothesisAgent ConditionsModification SiteOther...</div> <input type="text" value="Optional description (240 chars)"/> <input type="button" value="Submit"/>		
reach	...e of immune responses as well as differential regulation of different innate and adaptive ... has been implicated in the increased susceptibility of asthmatics to RV and in RV-	18234348	
reach	...a and exacerbations.		
reach	...th acute RV-induced asthma Wark et al., found that increased serum IP-10 levels but ... r IL-8 was specifically associated with infection and correlated with the degree of ... ction (Wark et al., 2007).	18234348	
reach	(b)A mouse model of RV-induced asthma exacerbation.	24278777	
reach	As RV is one the most common triggers of asthma exacerbations, it needs to be determined if blocking IL-4 and IL-13 could be useful in preventing experimental RV-induced exacerbation of asthma [24,50].	27088397	
reach	Some investigators have reported that RV-C caused more serious illness, especially wheezing and exacerbation of asthma, in some populations5, 9, 10, 11, 12, 13 compared with illnesses caused by RV -A and B.	31389049	

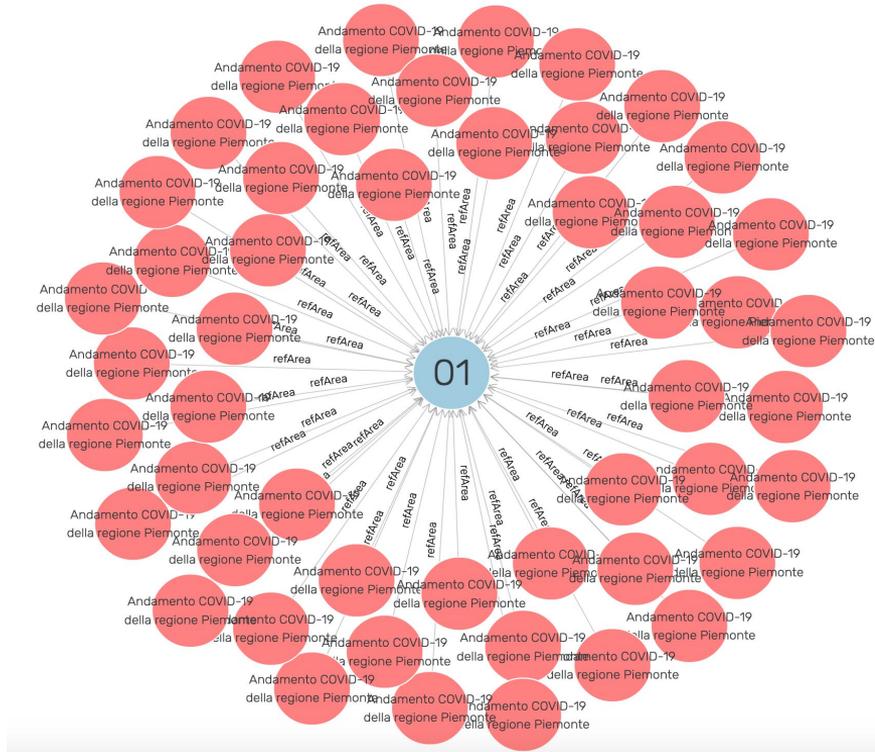
Building domain specific knowledge graphs

- We're collecting all possible COVID-19 data and archiving in our Dataverse
- Looking for various related controlled vocabularies and ontologies
- Building and reusing conversion pipelines to get all data values linked in RDF format

```
<http://www.protezionecivile.gov.it/dataset/covid19/national-trend/observations/20200224> a qb:Observation ;
  rdfs:label "Andamento nazionale di giorno 24/02/2020"^^xsd:string ;
  qb:dataset <http://www.protezionecivile.gov.it/dataset/covid19/national-trend> ;
  sdmx-dimension:refArea <https://w3id.org/italia/controlled-vocabulary/territorial-classifications/countries/italy/ITA> ;
  sdmx-dimension:refTime "2020-02-24"^^xsd:date ;
  dpc:deads "7"^^xsd:int ;
  dpc:healed "1"^^xsd:int ;
  dpc:homeIsolation "94"^^xsd:int ;
  dpc:hospitalizedWithSymptoms "101"^^xsd:int ;
  dpc:intensiveCare "26"^^xsd:int ;
  dpc:newPositive "221"^^xsd:int ;
  dpc:swabs "4324"^^xsd:int ;
  dpc:totalCases "229"^^xsd:int ;
  dpc:totalHospitalized "127"^^xsd:int ;
  dpc:totalPositive "221"^^xsd:int ;
  dpc:totalPositiveVariation "0"^^xsd:int .
```

The ultimate goal is to automate the process of the Knowledge extraction by using the latest developments in Artificial Intelligence and Deep Learning.

Visual graph of COVID-19 dataset



[Andamento COVID-19 della regione Piemonte](#)

Andamento COVID-19 della regione Piemonte

Types:

qb:Observation

RDF rank:

0

Search instance properties

sdmx-dimension:refTime

2020-03-12

dp:deads

26

dp:healed

0

dp:homelsolation

89

dp:hospitalizedWithSymptoms

368

dp:intensiveCare

97

dp:newPositive

79

dp:swabs

2879

Source: [CoronaWhy GraphDB](#)

SPARQL endpoint for CoronaWhy KG

Query 

<https://sparql.labs.coronawhy.org/sparql>

```
1 PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
2 PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
3 PREFIX mesh: <http://id.nlm.nih.gov/mesh/>
4 PREFIX grid: <https://www.grid.ac/institutes/>
5 PREFIX bf: <http://id.loc.gov/ontologies/bibframe/>
6 PREFIX loc: <http://id.loc.gov/ontologies/bibframe/source>
7 PREFIX mads: <http://www.loc.gov/mads/rdf/v1#>
8
9 select * from <http://coronawhy.org/> where
10 {
11 ?cord rdfs:label "United States" .
12 mesh:C0243052 ?a ?b
13 }
```

Table Response Gallery Chart Geo Geo-3D Geo events Pivot Timeline 10000 results in 3.051 seconds Filter query results Page size: 50

	cord	a	b
1	http://vufind.apps.coronawhy.org/vufind/Record/03bwk538#Place651-7	rdf:type	bf:Topic
2	http://vufind.apps.coronawhy.org/vufind/Record/094d0rn6#Place651-7	rdf:type	bf:Topic
3	http://vufind.apps.coronawhy.org/vufind/Record/2egeyh0j#Place651-7	rdf:type	bf:Topic
4	http://vufind.apps.coronawhy.org/vufind/Record/3p2rqavh#Place651-7	rdf:type	bf:Topic
5	http://vufind.apps.coronawhy.org/vufind/Record/5kapn32k#Place651-7	rdf:type	bf:Topic
6	http://vufind.apps.coronawhy.org/vufind/Record/6g55l35h#Place651-7	rdf:type	bf:Topic
7	http://vufind.apps.coronawhy.org/vufind/Record/7bkf7wrs#Place651-7	rdf:type	bf:Topic
8	http://vufind.apps.coronawhy.org/vufind/Record/7hcr406k#Place651-7	rdf:type	bf:Topic
9	http://vufind.apps.coronawhy.org/vufind/Record/8pd6jdb5#Place651-7	rdf:type	bf:Topic

Do you know a lot of people that can use SPARQL?

PubChemRDF Use Cases: SPARQL query



Q: What adverse effects of chemicals that are oral acute toxic according to GHS statement have been reported in PubMed literature, annotated by MeSH indexing?

```
PREFIX cito: <http://purl.org/spar/cito/>
PREFIX fabio: <http://purl.org/spar/fabio/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX sio: <http://semanticscience.org/resource/>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX meshv: <http://id.nlm.nih.gov/mesh/vocab#>
PREFIX mesh: <http://id.nlm.nih.gov/mesh/>

select distinct ?disease ?diseaselabel
where {
  ?compound sio:has-attribute/dcterms:subject/skos:broader/concept:Acute_Toxicity_Oral .
  ?syno sio:is-attribute-of ?compound .
  ?syno dcterms:subject ?meshconcept .
  ?pmid cito:discusses ?meshconcept .
  ?pmid fabio:hasSubjectTerm ?DQpair .
  ?DQpair meshv:hasQualifier mesh:Q000009 .
  ?pmid cito:discusses ?disease .
  ?disease rdf:type meshv:SCR_Disease .
  ?disease rdfs:label ?diseaselabel .
}
```

Source: [Semantic annotation](#) of the Laboratory Chemical Safety Summary in PubChem

CLARIAH conclusions



*“By developing these decentralised, yet controlled Knowledge Graph development practices we have contributed to increasing interoperability in the humanities and enabling new research opportunities to a wide range of scholars. However, **we observe that users without Semantic Web knowledge find these technologies hard to use, and place high value in end-user tools that enable engagement.** Therefore, for the future we emphasise the importance of tools to specifically target the goals of concrete communities – in our case, the analytical and quantitative answering of humanities research questions for humanities scholars. In this sense, usability is not just important in a tool context; in our view, we need to empower users in deciding under what models these tools operate.”* ([CLARIAH: Enabling Interoperability Between Humanities Disciplines with Ontologies](#))

Chicken-egg problem: users are building tools without data models and ontologies but in reality they need to build a knowledge graph with common ontologies first!

Linked Data integration challenges

- datasets are very heterogeneous and multilingual
- data usually lacks sufficient data quality control
- data providers using different modeling schemas and styles
- linked data cleansing and versioning is very difficult to track and maintain properly, web resources aren't persistent
- even modern data repositories providing only metadata records describing data without giving access to individual data items stored in files
- difficult to assign and manually keep up-to-date entity relationships in knowledge graph



World Wide Web Consortium

CoronaWhy has too much information streams that seems to be impossible to integrate and give back to COVID-19 researchers. So, do we have a solution?

Bibliographic Framework (BIBFRAME) as a Web of Data

“The Library of Congress officially launched its Bibliographic Framework Initiative in May 2011. The Initiative aims to re-envision and, in the long run, implement a new bibliographic environment for libraries that makes "the network" central and makes **interconnectedness commonplace**.”

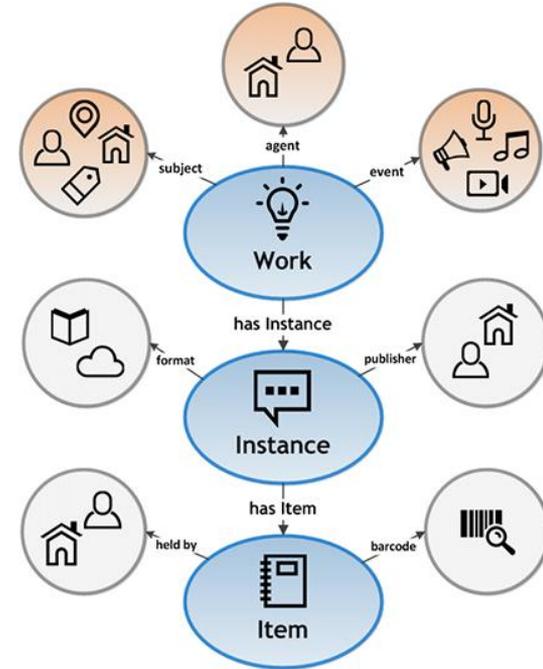
“Instead of thousands of catalogers repeatedly describing the same resources, the effort of one cataloger could be shared with many.” ([Source](#))

In 2019 BIBFRAME 2.0, the Library of Congress Pilot, was announced.

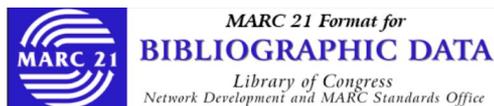
Let's take a journey and move from domain specific ontology to bibliographic!

BIBFRAME 2.0 concepts

- **Work.** The highest level of abstraction, a Work, in the BIBFRAME context, reflects the conceptual essence of the cataloged resource: authors, languages, and what it is about (subjects).
- **Instance.** A Work may have one or more individual, material embodiments, for example, a particular published form. These are Instances of the Work. An Instance reflects information such as its publisher, place and date of publication, and format.
- **Item.** An item is an actual copy (physical or electronic) of an Instance. It reflects information such as its location (physical or virtual), shelf mark, and barcode.
- **Agents:** Agents are people, organizations, jurisdictions, etc., associated with a Work or Instance through roles such as author, editor, artist, photographer, composer, illustrator, etc.
- **Subjects:** A Work might be “about” one or more concepts. Such a concept is said to be a “subject” of the Work. Concepts that may be subjects include topics, places, temporal expressions, events, works, instances, items, agents, etc.
- **Events:** Occurrences, the recording of which may be the content of a Work.



MARC as a foundation of the structured Data Hub



1999 Edition
Update No. 1 (October 2000) through Update No. 30 (May 2020)

This online publication provides access to both the full and concise versions of the *MARC 21 Format for Bibliographic Data*. The "full" bibliographic format contains detailed descriptions of every data element, along with examples, input conventions, and history sections. The "concise" bibliographic format contains abridged descriptions of every data element, along with examples. The full and concise versions are identified in the header of each field description.

Changes to the *MARC 21 Format for Bibliographic Data* that resulted from Update No. 30 (May 2020) are displayed in red print. The date located in the header of the full version of each field indicates the last month and year of update.

Table of Contents

- [Introduction](#) [Full | Concise]
- [Format Summary](#)
- [Leader](#) [Full | Concise]
- [Directory](#)
- [00X: Control Fields](#)
- [01X-09X: Numbers and Code Fields](#)
- [Heading Fields - General Information](#)
- [1XX: Main Entry Fields](#)
- [20X-24X: Title and Title-Related Fields](#)
- [25X-28X: Edition, Imprint, Etc. Fields](#)
- [3XX: Physical Description, Etc. Fields](#)
- [4XX: Series Statement Fields](#)
- [5XX: Note Fields](#)
- [6XX: Subject Access Fields](#)
- [70X-75X: Added Entry Fields](#)
- [76X-78X: Linking Entry Fields](#)
- [80X-83X: Series Added Entry Fields](#)
- [841-88X: Holdings, Location, Alternate Graphics, Etc. Fields](#)
- [Appendix A: Control Subfields](#)
- [Appendix B: Full Level Record Examples](#)
- [Appendix C: Minimal Level Record Examples](#)
- [Appendix D: Multiscript Records](#)
- [Appendix E: Alphabetical List of Ambiguous Headings](#)
- [Appendix F: Initial Definite and Indefinite Articles](#)
- [Appendix G: Format Changes for Update No. 30 \(May 2020\)](#)
- [Appendix H: Local Data Elements](#)

MARC standard was developed in the 1960s to create records that could be read by computers and shared among libraries. The term MARC is an abbreviation for MACHine Readable Cataloging.

The **MARC 21** bibliographic format was created for the international community. It's very rich, with **more** than 2,000 data elements defined!

It's identified by its ISO (International Standards Organization) number: ISO 2709

Source: [the Library of Congress](#), USA

How to integrate data in the common KG?

- Use MARC 21 as a basis for all bibliographic and authority records
- All controlled vocabularies should be expressed in MARC 21 format for Authority Data, we need to build an authority linking process with the “human in the loop” approach that will allow to verify AI predicted links.
- Different MARC 21 fields could be linked to the different ontologies and/or even interlinked. For example, we can get some entities linked to both MeSH and Wikidata in the same bibliographic record to increase the interoperability of the Knowledge Graph.
- Every COVID-19 paper can get a metadata enrichment provided by any research team working on the NLP extraction of entities, relations or linking CV together.

COVID-19 paper in MARC 21 representation

Holdings	Description	Comments	Similar Items	Staff View
LEADER	06032cas a2201165	4500		
001	o32e9o0l			
008	cas cc chil			
245	1	0	ja The Transformation of Enterovirus Replication Structures: a Three-Dimensional Study of Single- and Double-Membrane Compartments	
856	4	0	ju https://doi.org/10.1128/mbio.00166-11	
651			jx coronavirus	
651	4		ja Netherlands	
370			je Leiden, Netherlands	
852			ja Department of Medical Microbiology, Molecular Virology Laboratory, Center of Infectious Diseases, Leiden University Medical Center, Leiden, The Netherlands	
650	1	2	ja Act Relationship Type - transformation x C1554215	
650	1	2	ja Three-dimensional x C0450363	
650	1	2	ja Induce (action) x C0205263	
650	1	2	ja Membrane x C0596901	
650	1	2	ja Cells x C0007634	
650	1	2	ja Cell Microenvironment x C1707328	
650	1	2	ja RNA, Viral x C0035736	
650	1	2	ja Anabolism x C0220781	
650	1	2	ja structure x C0678594	
650	1	2	ja Enterovirus x C0014383	
650	1	2	ja Family Picornaviridae x C0031886	
650	1	2	ja Tissue membrane x C0025255	
650	1	2	ja Architecture as Topic x C0003737	

Structure of the bibliographic record:

- authority records contain information about authors and affiliation
- Medical entities extracted by NLP pipeline interlinked in 650x fields
- part of metadata fields generated and filled by Machine Learning models, part contributed by human experts
- provenance information kept in 833x fields indicating fully or partially machine-generated records
- Relations between entities stored in 730x fields

Vufind discovery tool for libraries powered by MARC 21

The screenshot displays the CoronaWhy VuFind search interface. At the top, the logo 'CORONAWHY' is on the left, and navigation links for 'Home', 'Daily Progress', 'Solutions', 'Data', and 'Team' are on the right, along with a 'JOIN THE FIGHT!' button. Below the navigation is a search bar with a 'Find' button and a 'Language: English' dropdown. The search results section shows 'Suggested Topics within your search' with filters for 'coronavirus' (203,451), 'Medicine' (172,412), 'COVID-19' (133,230), 'C1880229' (4,507), and 'DICOM Study' (4,507). The main results list shows 1-20 results of 203,461 for search ' '. The first result is 'OnabotulinumtoxinA infiltration and nerve blocks in patients with headache and neuralgia: safety recommendations to prevent SARS-CoV-2 infection' by Santos-Lasaosa, S., Porta-Etessam, J., published in 2020. The right sidebar contains a 'Narrow Search' section with filters for Institution (203,461), Library (203,461), Format (Serial: 133,229, Book: 70,222, Journal: 10, Microfilm: 3), Call Number (A - General Works: 10), and Author (CoronaWhy Labs Labs: 102,036, University of Hong Kong: 744).

Source: [CoronaWhy VuFind](#)

Landing page of publications from CORON-19

Home Daily Progress Solutions Data Team [JOIN THE FIGHT!](#)

All Fields [Advanced](#)

[Reset Filters](#) Language: English Suggested Topics: C1880229

[Search](#) / [The Transformation of Enterovi...](#) / Holdings

[Cite this](#) [Text this](#) [Email this](#) [Print](#) [Export Record](#) [Save to List](#)



The Transformation of Enterovirus Replication Structures: a Three-Dimensional Study of Single- and Double-Membrane Compartments

All positive-strand RNA viruses induce membrane structures in their host cells which are thought to serve as suitable microenvironments for viral RNA synthesis. The structures induced by enteroviruses, which are members of the family Picornaviridae, have so far been described as either single- or do...

[Full description](#)

Main Authors:	Limpens, Ronald W. A. L. , van der Schaar, Hilde M. , Kumar, Darshan , Koster, Abraham J. , Snijder, Eric J. , van Kuppeveld, Frank J. M. , Bárcena, Montserrat
Corporate Authors:	Leiden University Medical Center , Radboud University Nijmegen Medical Centre
Format:	Serial
Language:	English
Published:	mBio 2011
Subjects:	Act Relationship Type - transformation > C1554215 Three-dimensional > C0450363 Induce (action) > C0205263 Membrane > C0596901 Cells > C0007634 Cell Microenvironment > C1707328 RNA, Viral > C0035736 Anabolism > C0220781

Similar Items

- [SARS-Coronavirus Replication Is Supported by a Reticulovesicular Network of Modified Endoplasmic Reticulum](#)
by: Knoops, Kévin, et al.
Published: (2008)
- [Three-Dimensional Architecture and Biogenesis of Membrane Structures Associated with Hepatitis C Virus Replication](#)
by: Romero-Brey, Inés, et al.
Published: (2012)
- [Packaging of Genomic RNA in Positive-Sense Single-Stranded RNA Viruses: A Complex Story](#)
by: Comas-Garcia, Mauricio
Published: (2019)
- [Role of Cellular Lipids in Positive-Sense RNA Virus Replication Complex Assembly and Function](#)
by: Stapleford, Kenneth A., et al.
Published: (2010)

CORD-19 collection in BIBFRAME 2.0

```
43 <bf:Content rdf:about="http://id.loc.gov/vocabulary/contentTypes/txt">
44 <rdfs:label>text</rdfs:label>
45 </bf:Content>
46 </bf:content>
47 <bf:language>
48 <bf:Language rdf:about="http://id.loc.gov/vocabulary/languages/%2Fch"/>
49 </bf:language>
50 <rdfs:label>The SARS Coronavirus S Glycoprotein Receptor Binding Domain: Fine Mapping and Functional Characterization</rdfs:label>
51 <bf:title>
52 <rdfs:label>The SARS Coronavirus S Glycoprotein Receptor Binding Domain: Fine Mapping and Functional Characterization</rdfs:label>
53 <bf:titleSortKey>The SARS Coronavirus S Glycoprotein Receptor Binding Domain: Fine Mapping and Functional Characterization</bf:titleSortKey>
54 <bf:mainTitle>The SARS Coronavirus S Glycoprotein Receptor Binding Domain: Fine Mapping and Functional Characterization</bf:mainTitle>
55 </bf:title>
56 </bf:Title>
57 </bf:subject>
58 <bf:subject>
59 <bf:Place rdf:about="http://vufind.apps.coronawhy.org/vufind/Record/ofx0hvvs#Place651-6">
60 <rdf:type rdf:resource="http://www.loc.gov/mads/rdf/v1#ComplexSubject"/>
61 <rdfs:label>coronavirus</rdfs:label>
62 <madsrdf:authoritativeLabel>coronavirus</madsrdf:authoritativeLabel>
63 <madsrdf:componentList rdf:parseType="Collection">
64 <madsrdf:Topic>
65 <madsrdf:authoritativeLabel>coronavirus</madsrdf:authoritativeLabel>
66 </madsrdf:Topic>
67 </madsrdf:componentList>
68 </bf:Place>
69 </bf:subject>
70 <bf:subject>
71 <bf:Place rdf:about="http://vufind.apps.coronawhy.org/vufind/Record/ofx0hvvs#Place651-7">
72 <rdf:type rdf:resource="http://www.loc.gov/mads/rdf/v1#Geographic"/>
73 <rdfs:label>United States</rdfs:label>
74 <madsrdf:authoritativeLabel>United States</madsrdf:authoritativeLabel>
75 </bf:Place>
76 </bf:subject>
77 <bf:subject>
78 <bf:Topic rdf:about="http://id.nlm.nih.gov/mesh/C1175175">
79 <rdf:type rdf:resource="http://www.loc.gov/mads/rdf/v1#ComplexSubject"/>
80 <rdfs:label>Severe Acute Respiratory Syndrome--C1175175</rdfs:label>
81 <madsrdf:authoritativeLabel>Severe Acute Respiratory Syndrome--C1175175</madsrdf:authoritativeLabel>
82 <madsrdf:componentList rdf:parseType="Collection">
83 <madsrdf:Topic>
84 <madsrdf:authoritativeLabel>Severe Acute Respiratory Syndrome</madsrdf:authoritativeLabel>
85 </madsrdf:Topic>
86 <madsrdf:Topic>
87 <madsrdf:authoritativeLabel>C1175175</madsrdf:authoritativeLabel>
88 </madsrdf:Topic>
89 </madsrdf:componentList>
90 <bf:source>
91 <bf:Source>
92 <bf:code>mesh</bf:code>
93 </bf:Source>
94 </bf:source>
95 </bf:Topic>
96 </bf:subject>
97 <bf:subject>
98 <bf:Topic rdf:about="http://id.nlm.nih.gov/mesh/C1283195">
99 <rdf:type rdf:resource="http://www.loc.gov/mads/rdf/v1#ComplexSubject"/>
100 <rdfs:label>Mapping (action)--C1283195</rdfs:label>
101 <madsrdf:authoritativeLabel>Mapping (action)--C1283195</madsrdf:authoritativeLabel>
102 <madsrdf:componentList rdf:parseType="Collection">
103 <madsrdf:Topic>
104 <madsrdf:authoritativeLabel>Mapping (action)</madsrdf:authoritativeLabel>
105 </madsrdf:Topic>
```

CoronaWhy Graph published as RDF

 Search ▾ User Guide Support Sign Up Log In

 COVID-19 Knowledge Graphs (CoronaWhy)

COVID-19 Data Hub > COVID-19 Knowledge Graphs >

CORD-19 collection in RDF using MARC21 ontology

Version 5.3

 Tykhonov, Vyacheslav, 2020, "CORD-19 collection in RDF using MARC21 ontology", <https://doi.org/10.5072/FK2/BO5DLV>, COVID-19 Data Hub, V5

Cite Dataset ▾ Learn about Data Citation Standards.

Access Dataset ▾
Contact Owner Share

Dataset Metrics ⓘ
345 Downloads ⓘ

Description ⓘ

A sample from CORD-19 collection published in MARC21 standard and exposed as Resource Description Framework (RDF) for testing purposes. It's a first step to create a Knowledge Graph out of all COVID-19 papers. Sample data linked to:

- Medical Subject Headings (MeSH) thesaurus
- GRID database of affiliations
- Geonames database of locations

Download all files and put in /tmp/data folder, use isql-v to upload this RDF collection and query it in Virtuoso:

```
id_dir('/tmp/data','*.rdf','http://coronawhy.org/');  
rdf_loader_run (log_enable=>3);  
checkpoint;
```

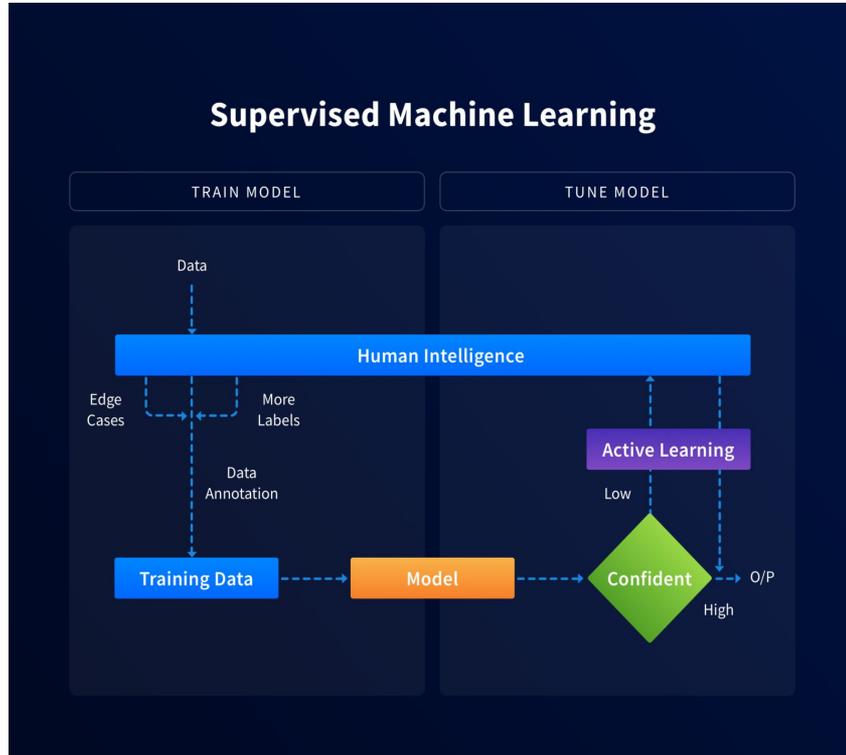
Sample SPARQL query to access CoronaWhy COVID-19 Graph:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>  
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>  
PREFIX mesh: <http://id.nlm.nih.gov/mesh/>  
PREFIX grid: <https://www.grid.ac/institutes/>  
PREFIX bf: <http://id.loc.gov/ontologies/bibframe/>  
PREFIX loc: <http://id.loc.gov/ontologies/bibframe/source>  
PREFIX mads: <http://www.loc.gov/mads/rdf/v1#>
```

```
select * from <http://coronawhy.org/> where  
{  
?cord rdfs:label "United States".
```

Source: [CoronaWhy Dataverse](https://dataverse.org/dataset.xhtml?persistentId=doi:10.5072/FK2/BO5DLV)

Human-in-the-Loop for Machine Learning



“Computers are incredibly fast, accurate and stupid; humans are incredibly slow, inaccurate and brilliant; together they are powerful beyond imagination.”

Albert Einstein

“A combination of AI and Human Intelligence gives rise to an extremely high level of accuracy and intelligence (Super Intelligence)”

CLARIAH and Network Digital Heritage (NDE) GraphiQL

```
GraphiQL History
1 { terms(match:"COVID",dataset:["wikidata"])
2   { dataset terms {uri, altLabel} }
3
4 }
```

```
{
  "data": {
    "terms": [
      {
        "dataset": "wikidata",
        "terms": [
          {
            "uri": "http://www.wikidata.org/entity/Q18975243",
            "altLabel": [
              "coronavirus disease"
            ]
          },
          {
            "uri": "http://www.wikidata.org/entity/Q84263196",
            "altLabel": [
              "COVID-19"
            ]
          },
          {
            "uri": "http://www.wikidata.org/entity/Q81068910",
            "altLabel": [
              "COVID-19 pandemic"
            ]
          },
          {
            "uri": "http://www.wikidata.org/entity/Q82069695",
            "altLabel": [
              "SARS-CoV-2"
            ]
          }
        ]
      }
    ]
  }
}
```

```
<?xml version="1.0" encoding="UTF-8"?>
<nde:NDE xmlns:nde="https://www.networkdigitalerfgoed.nl/">
  <nde:dataset id="clavas" recipe="nl.knaw.huc.di.nde.recipe.OpenSK05">
    <nde:label xml:lang="nl">ISO 639-3 taalcodes (CLAVAS)</nde:label>
    <nde:api>https://clavas.clarin.eu/clavas/public/api</nde:api>
    <nde:conceptScheme>http://hdl.handle.net/11459/CLAVAS_810f8d2a-6723-3ba6-2e57-41d6d3844816</nde:conceptScheme>
  </nde:dataset>
  <nde:dataset id="wikidata" recipe="nl.knaw.huc.di.nde.recipe.WikiData">
    <nde:label xml:lang="nl">Wikidata</nde:label>
    <nde:api>https://www.wikidata.org</nde:api>
    <nde:wildcard>no</nde:wildcard>
  </nde:dataset>
  <nde:dataset id="wikidatagtaa" recipe="nl.knaw.huc.di.nde.recipe.WikiDataGTAAConcepts">
    <nde:label xml:lang="nl">Wikidata/GTAA: wikidata entities that are linked to the GTAA</nde:label>
    <nde:api>https://www.wikidata.org</nde:api>
  </nde:dataset>
  <nde:dataset id="gtaa" recipe="nl.knaw.huc.di.nde.recipe.OpenSK05">
    <nde:label xml:lang="nl">GTAA (B&G)</nde:label>
    <nde:api>http://openskos.beeldengeluid.nl/api</nde:api>
    <nde:tenant>beng</nde:tenant>
    <nde:collection>gtaa</nde:collection>
  </nde:dataset>
  <nde:dataset id="gtaaonderwerpen" recipe="nl.knaw.huc.di.nde.recipe.OpenSK05">
    <nde:label xml:lang="nl">GTAA Onderwerpen (B&G)</nde:label>
    <nde:api>http://openskos.beeldengeluid.nl/api</nde:api>
    <nde:conceptScheme>http://data.beeldengeluid.nl/gtaa/Onderwerpen</nde:conceptScheme>
  </nde:dataset>
</nde:NDE>
```

CoronaWhy is running own instance of [NDE on Kubernetes](#) cluster and maintains the support of another ontologies (MeSH, ICD, NDF) available via their SPARQL endpoints!

If you want to query API:

```
curl
```

```
"http://nde.dataverse-dev.coronawhy.org/nde/graphql?query=%20%2B%20terms(match%3A%22COVID%22%2Cdataset%3A%5B%22wikidata%22%5D)%20%2B%20dataset%20terms%20%7Buri%2C%20altLabel%7D%20%7D%20%7D"
```

Increasing Dataverse metadata interoperability

Subject * ?

Keyword ?

Vocabulary ?

wikidata

Term ?

COVID-19



Vocabulary URL ?

http://www.wikidata.org/entity/Q810689

Vocabulary ?

wikidata

Term ?

covi



Vocabulary URL ?

- COVID-19
- COVID-19 pandemic
- COVID-19 pandemic in India
- Covi
- Covi
- Covi
- SARS-CoV-2



Related Publication ?

Citation ?

Hypothes.is annotations as a peer review service

The screenshot shows a Hypothes.is interface. At the top left is the CORONA WHY logo. The navigation bar includes 'Home', 'Hypothesis Test', 'Daily Progress', and 'Calendar'. The main text is titled 'Ensuring global access to COVID-19 vaccines'. The text contains several paragraphs with yellow highlights. On the right side, there is a vertical list of annotations. Each annotation is by user 'cchen1111' and is public. The annotations include:

- An annotation on the first paragraph: 'However, models developed by the Imperial College COVID-19 Response Team suggest that "transmission will quickly rebound if interventions are relaxed".' The highlighted text is 'transmission will quickly rebound if interventions are relaxed'.
- An annotation on the second paragraph: 'the World Bank'.
- An annotation on the third paragraph: '12-18 months'.
- An annotation on the fourth paragraph: 'US\$2 billion'.

1. AI pipeline does domain specific entities extraction and ranking of relevant COVID-19 papers.
2. Automatic entities and statements will be added, important fragments should be highlighted.
3. Human annotators should verify results and validate all statements.

Doccano annotation with Machine Learning

The screenshot displays the Doccano web interface. On the left, a sidebar contains a search bar, a 'Sort by' dropdown, and a list of results. The main area shows a document with various terms highlighted in colored boxes, each with a small icon representing a category. A legend at the top of the document lists these categories: AMINO_ACID (a), CELL (c), CELLULAR_COMPONENT (C-c), CELL_LINE (S-c), CELL_TYPE (C-S-c), CHEBI (h), CHEMICAL (C-h), CL (l), DISEASE (d), DNA (C-d), GENE_OR_GENE_PRODUCT (g), GGP (C-g), GO (S-g), ORGANISM (o), PROTEIN (p), RNA (r), SIMPLE_CHEMICAL (s), SO (C-s), TAXON (t), and UMLS (u).

An atypical RNA pseudoknot stimulator and an upstream attenuation signal for -1 ribosomal -1 ribosomal frameshifting of SARS SARS coronavirus. The -1 ribosomal frameshifting requires the existence of an in cis RNA slippery sequence cis RNA RNA slippery sequence and is promoted by a downstream stimulator RNA. An atypical RNA pseudoknot RNA pseudoknot pseudoknot with an extra stem formed by complementary sequences complementary sequences within loop 2 of an H-type pseudoknot is characterized in the severe acute respiratory syndrome coronavirus coronavirus (SARS CoV) genome. This pseudoknot can serve as an efficient stimulator for -1 -1 frameshifting in vitro. Mutational analysis of the extra stem suggests frameshift efficiency can be modulated via manipulation of the

[Source: Doccano Labs](#)

Building an Operating System for Open Science

CoronaWhy Common Research and Data Infrastructure is distributed and robust enough to be scaled up and reused for other tasks like cancer research

All services are build from Open Source components

Data processed and published in FAIR way, the provenance information is the part of our Data Lake

Data evaluation and credibility is the top priority, we're providing tools for the expert community for the verification of our datasets

The transparency of data and services guarantees the reproducibility of all experiments and get bring new insights in COVID-19 research

CoronaWhy Common Research and Data Infrastructure

[Data preprocessing pipeline](#) implemented on Jupyter notebook Docker with extra modules.

[Dataverse](#) as data repository to store data from automatic and curated workflows.

[Elasticsearch](#) has COVID-19 indexes on sections and sentences level with spacy enrichments. Other indexes: MeSH, Geonames, GRID.

[Hypothesis](#) and Doccano annotation services to annotate publications

[Virtuoso](#) and GraphDB with public SPARQL endpoints to query COVID-19 Knowledge Graph

Other services: [Colab](#), [MongoDB](#), [Kibana](#), [BEL Commons 3.0](#), INDRA, [Geoparser](#), Tabula

<https://github.com/CoronaWhy/covid-19-infrastructure>

CoronaWhy Data API 0.1 OAS3

/openapi.json

CoronaWhy is globally distributed, volunteer-powered research organisation. We're trying to assist the medical community's ability to answer key questions related to COVID-19.

country

Put this citation in working papers and published papers that use this dataset: Guidotti et al., (2020). COVID-19 Data Hub. Journal of Open Source Software, 5(51), 2376, <https://doi.org/10.21105/joss.02376>

GET /country/{item_id} Data Item

dataverse

Dataverse integration by API. Available actions: [showfiles, getfile]

GET /dataverse/{action} Dataverse

cord

Metadata by cord_id

CORD-19 collection access by cord_id

GET /cord/ Read Cord

altmetrics

Altmetrics by DOI or cord_id

CORD-19 papers Altmetrics

GET /altmetrics/ Read Altmetrics

search

CORD-19 search

CORD-19 papers search

GET /cordsearch/ Search Cord

Source: [CoronaWhy API](#) built on FastAPI framework for Python

Thank you! Questions?

@Slava Tykhonov on CoronaWhy Slack

vyacheslav.tykhonov@dans.knaw.nl

www.coronawhy.org

www.dans.knaw.nl/en